Managing extreme AI risks amid rapid progress

Year: 2023 | Citations: 296 | Authors: Y. Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel

---

Abstract

Preparation requires technical research and development, as well as adaptive, proactive governance Artificial intelligence (AI) is progressing rapidly, and companies are shifting their focus to developing generalist AI systems that can autonomously act and pursue goals. Increases in capabilities and autonomy may soon massively amplify AI’s impact, with risks that include large-scale social harms, malicious uses, and an irreversible loss of human control over autonomous AI systems. Although researchers have warned of extreme risks from AI (1), there is a lack of consensus about how to manage them. Society’s response, despite promising first steps, is incommensurate with the possibility of rapid, transformative progress that is expected by many experts. AI safety research is lagging. Present governance initiatives lack the mechanisms and institutions to prevent misuse and recklessness and barely address autonomous systems. Drawing on lessons learned from other safety-critical technologies, we outline a comprehensive plan that combines technical research and development (R&D) with proactive, adaptive governance mechanisms for a more commensurate preparation.